IQLA
International Quantitative Linguistics Association

QUALICO 2025
Brno, Czech Republic
June 26-28, 2025

Quantitative Linguistics and Large Language Models: Epistemological Challenges and Applications

Antoni Hernández-Fernández*, Bernardino Casas, Jaume Baixeries, Neus Català

antonio.hernandez@upc.edu @therfer

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

INSTITVT D'ESTVDIS CATALANS MCMVII Institut d'Estudis Catalans

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

FUNDING

Semàntica de les paraules del català: teoria i aplicacions clíniques PRO2022-S03. **Secció de Ciències i Tecnologia, Institut d'Estudis Catalans.** https://futur.upc.edu/32847767

Institut de Ciències de l'Educació
UNIVERSITAT POLITÈCNICA DE CATALUNYA

antonio.hernandez@upc.edu @therfer

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

Initial clarification

What follows is intended to be nothing more than an excuse for debate...

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

SUMMARY

Exposition: LLM and QL

Rising action: conflict and climax

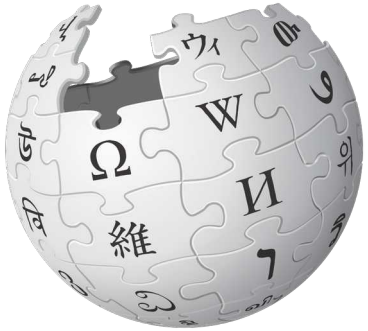
Falling action

Discussion

Large Language Models (LLM)

LLM is **a type of machine learning model designed for natural language processing (NLP)** tasks such as language generation.

LLMs are ***language models*** with many parameters, and are trained with self-supervised learning on a vast amount of text. The largest and most capable LLMs are **generative pretrained transformers (GPTs)**.

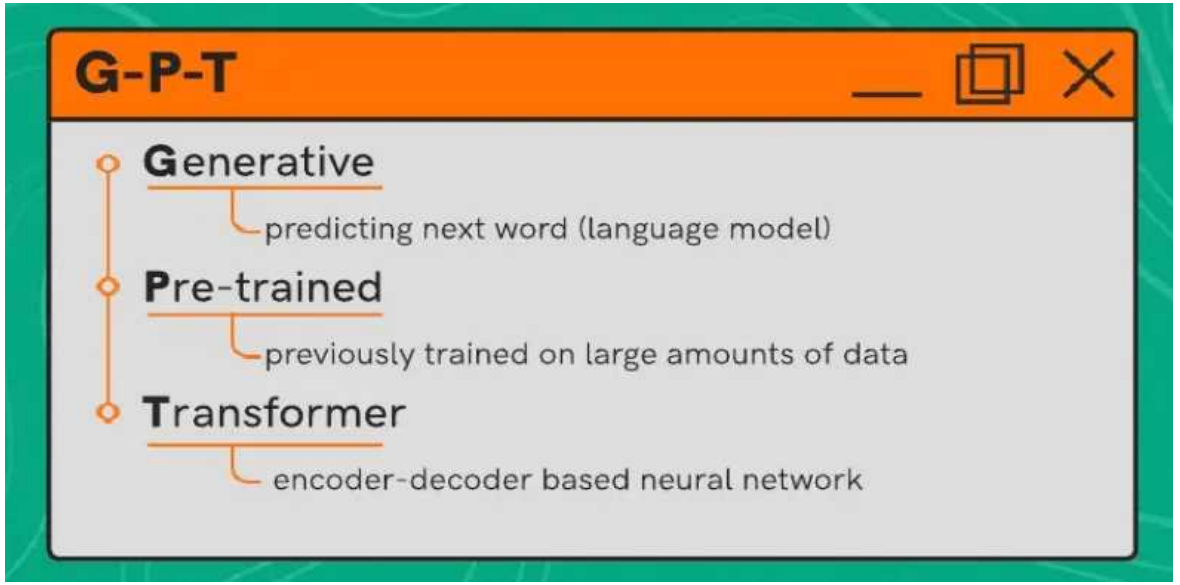


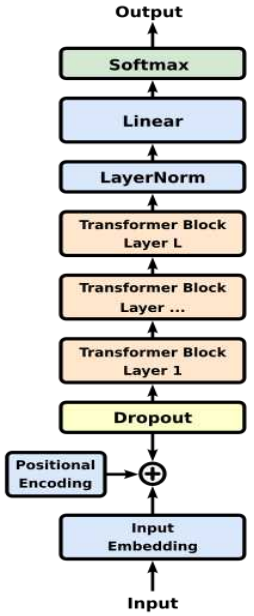
These **models acquire predictive power regarding syntax, semantics, and ontologies inherent in human language corpora**, but they also inherit inaccuracies and biases present in the data they are trained in.

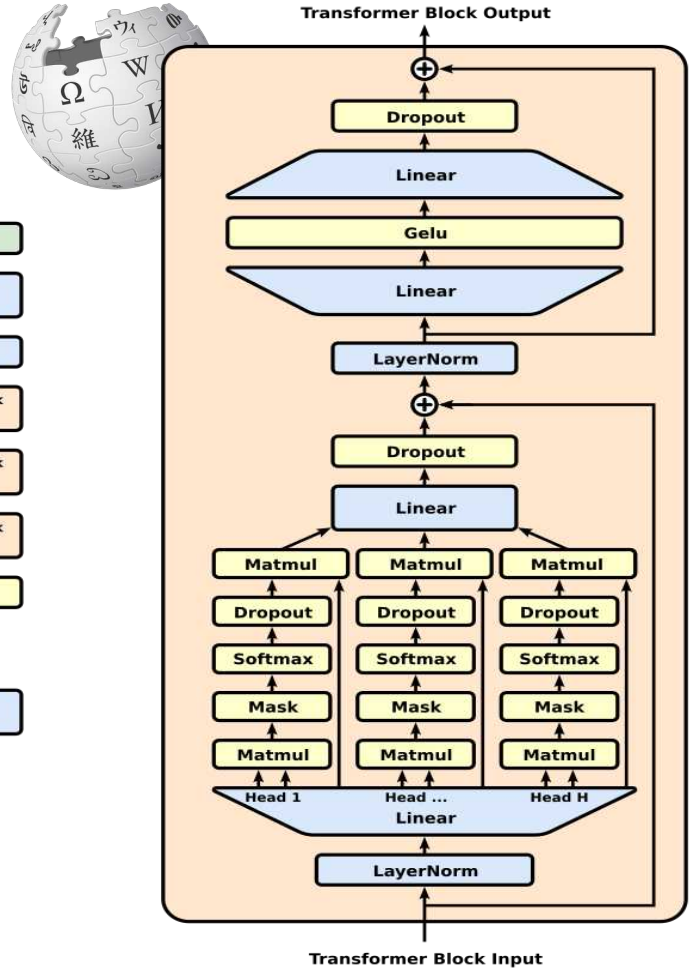
https://en.wikipedia.org/wiki/Large_language_model

Generative Pre-Trained Transformer (GPT)

Based on the **transformer deep learning architecture** (Vaswani et al., 2017), **pre-trained on large data sets** of unlabeled text, and **able to generate novel human-like content**.







https://en.wikipedia.org/wiki/Generative_pre-trained_transformer
https://oeiac.cat/en/

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

Generative Pre-Trained Transformer (GPT)

Exposition

Rising action

Falling action

Discussion

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Google Research Google Brain

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com

Karthik Narasimhan OpenAI karthikn@openai.com

Tim Salimans OpenAI tim@openai.com

Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

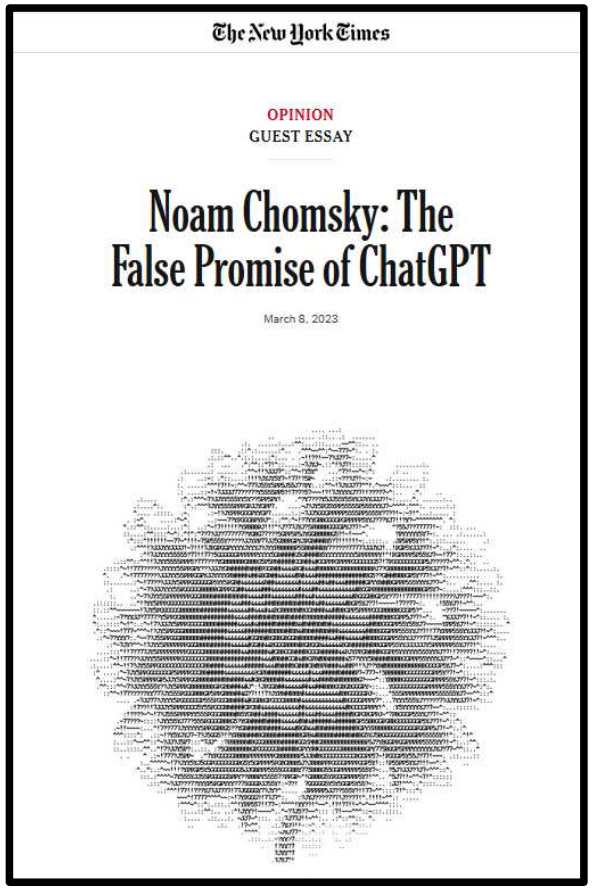
OpenAI

ChatGPT

Vaswani et al, 2017 https://arxiv.org/pdf/1706.03762

Radford et al, 2018 https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Linguistic debate: Large Language Models

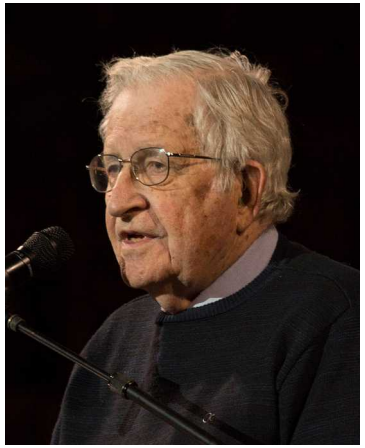

The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations.



Chomsky et al. (2023) https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Linguistic debate: Large Language Models

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

Modern machine learning has subverted and bypassed the theoretical framework of Chomsky's generative approach to linguistics, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine *theories* of language, and I highlight the links between these models and approaches to linguistics that are based on gradient computations and memorized constructions. I also describe why these models undermine strong claims for the innateness of language and respond to several critiques of large language models, including arguments that they can't answer "why" questions and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

i) LLM implement genuine *theories* of language

ii) LLM undermine innateness

iii) LLM discover grammar

Piantadosi (2023) https://zenodo.org/records/12665933

https://slator.com/how-large-language-models-prove-chomsky-wrong-with-steven-piantadosi/

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

LLM, only engineering achievement?

Exposition

Rising action

Falling action

Discussion

Language Sciences 106 (2024) 101672

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Large models of what? Mistaking engineering achievements for human linguistic agency

Abeba Birhane [a], Marek McGann [b,*]

[a] Mozilla Foundation and School of Computer Science and Statistics, Trinity College Dublin, Ireland
[b] Department of Psychology, Mary Immaculate College, Limerick, Ireland

ARTICLE INFO

Article history:
Available online 21 August 2024

Keywords:
Large language models
Algospeak
Agency
Precariousness
Embodiment
Language
Enaction
Precarity

ABSTRACT

In this paper we argue that key, often sensational and misleading, claims regarding linguistic capabilities of Large Language Models (LLMs) are based on at least two unfounded assumptions: the *assumption of language completeness* and the *assumption of data completeness*. Language completeness assumes that a distinct and complete thing such as "a natural language" exists, the essential characteristics of which can be effectively and comprehensively modelled by an LLM. The assumption of data completeness relies on the belief that a language can be quantified and wholly captured by data. Work within the enactive approach to cognitive science makes clear that, rather than a distinct and complete thing, language is a means or way of acting. Languaging is not the kind of thing that can admit of a complete or comprehensive modelling. From an enactive perspective we identify three key characteristics of enacted language; *embodiment*, *participation*, and *precariousness*, that are absent in LLMs, and likely incompatible in principle with current architectures. We argue that these absences imply that LLMs are not now and cannot in their present form be linguistic agents the way humans are. We illustrate the point in particular through the phenomenon of "algospeak", a recently described pattern of high-stakes human language activity in heavily controlled online environments. On the basis of these points, we conclude that sensational and misleading claims about LLM agency and capabilities emerge from a deep misconception of both what human language is and what LLMs are.

© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Birhane & McGann (2024)

nature

Explore content · About the journal · Publish with us · Subscribe

nature > correspondence > article

CORRESPONDENCE | 19 March 2024

Three reasons why AI doesn't model human language

By Johan J. Bolhuis, Stephen Crain, Sandiway Fong & Andrea Moro

Artificial intelligence (AI) is being used to develop large language models (LLMs) with considerable success. But they should not be seen as being models of how human language works and is acquired.

First, LLMs are probabilistic models of externalized language data, whereas human language is truly generative: it yields an unbounded number of hierarchically structured expressions (M. B. A. Everaert *et al. Trends Cogn. Sci.* **19**, 729–743; 2015). Second, language acquisition in infants does not depend on massive amounts of input data, but includes knowledge of language's generative nature. Therefore, children can acquire any language rapidly with minimal linguistic input (C. Yang *et al. Neurosci. Biobehav. Rev.* **81**, 103–119; 2017). Third, LLMs can produce 'impossible' languages, not generated by the principles governing all known human languages, just as well as (if not better than) natural language output, and cannot distinguish between them (A. Moro *et al. Cortex* **167**, 82–85; 2023).

LLMs are a useful technology, good at reproducing strands of information derived from training data, but they do not model the structure of human language, let alone encompass its creative nature.

Nature **627**, 489 (2024)

doi: https://doi.org/10.1038/d41586-024-00824-z

Bolhuis et al. (2024)

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

Do LLMs make human-like linguistic generalizations?

Exposition

Rising action

Falling action

Discussion

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias

Vittoria Dentella [a,1], Fritz Günther [b], and Evelina Leivada [c,d]

Edited by Susan Goldin-Meadow, University of Chicago, Chicago, IL; received June 7, 2023; accepted October 28, 2023

December 13, 2023 | 120 (51) e2309583120 | https://doi.org/10.1073/pnas.2309583120

Significance

The synthetic language generated by recent Large Language Models (LMs) strongly resembles the natural languages of humans. This resemblance has given rise to claims that LMs can serve as the basis of a theory of human language. Given the absence of transparency as to what drives the performance of LMs, the characteristics of their language competence remain vague. Through systematic testing, we demonstrate that LMs perform nearly at chance in some language judgment tasks, while revealing a stark absence of response stability and a bias toward yes-responses. Our results raise the question of how knowledge of language in LMs is engineered to have specific characteristics that are absent from human performance.

Dentella, Günther & Leivada (2023)

PNAS LETTER

Language models align with human judgments on key grammatical constructions

Jennifer Hu[a,1], Kyle Mahowald[b], Gary Lupyan[c], Anna Ivanova[d], and Roger Levy[e]

Affiliations are included on p. 3.

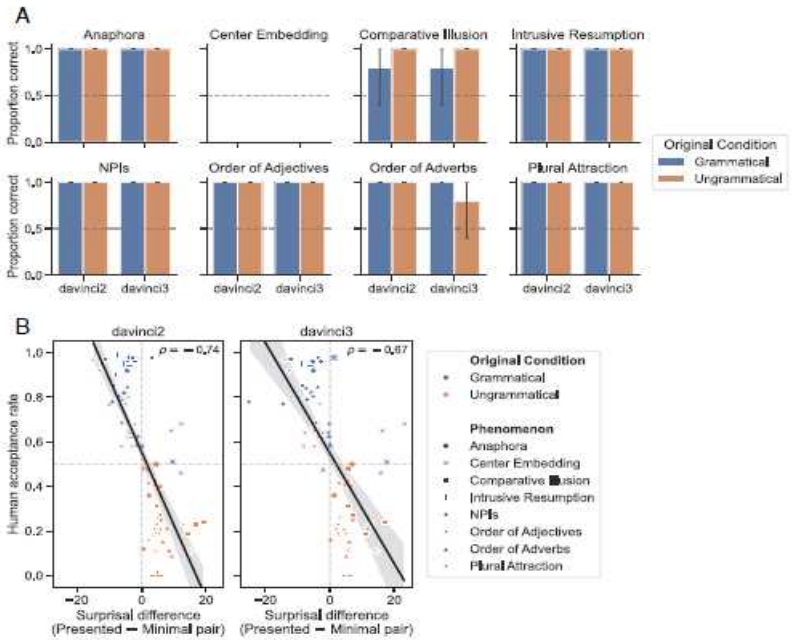


Fig. 1. (A) Accuracy scores achieved by models on a version of DGL's original materials with minimal pairs. For each phenomenon, accuracy is computed as the proportion of items in that phenomenon where the model assigns higher probability to the grammatical version of that item (minimal pair) than the ungrammatical version. (B) x-axis: Difference in sum surprisal (negative log probability) between the sentence presented to humans in DGL's experiments versus its counterpart in the minimal pair. y-axis: Human acceptance rate (proportion judged as grammatical) for the presented sentence in each minimal pair. Each point represents a minimal pair test item.

Hu et al. (2024)

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

Vote!

Exposition

Rising action

Falling action

Discussion

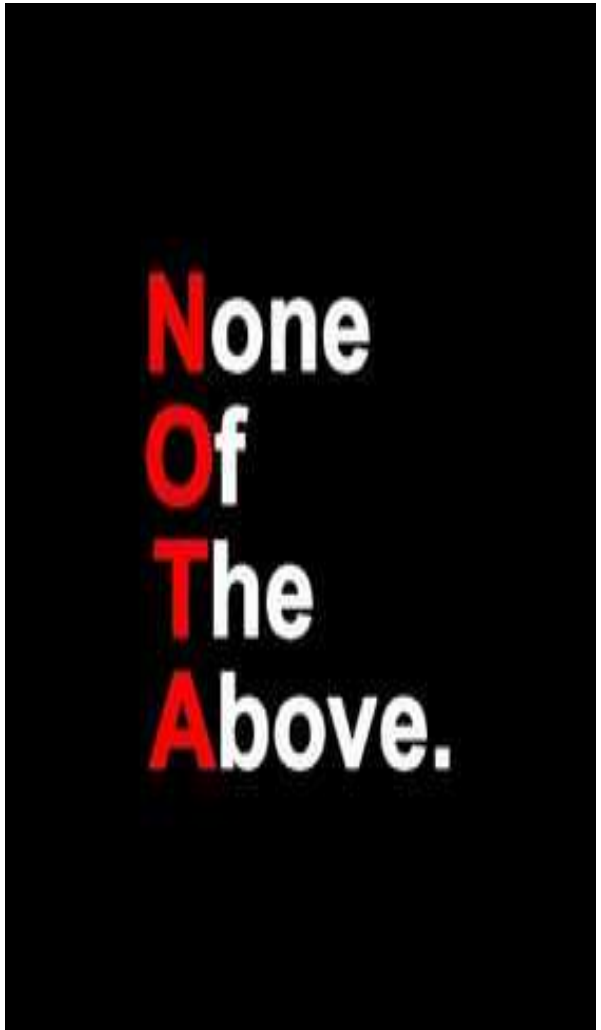
Institut d'Estudis Catalans

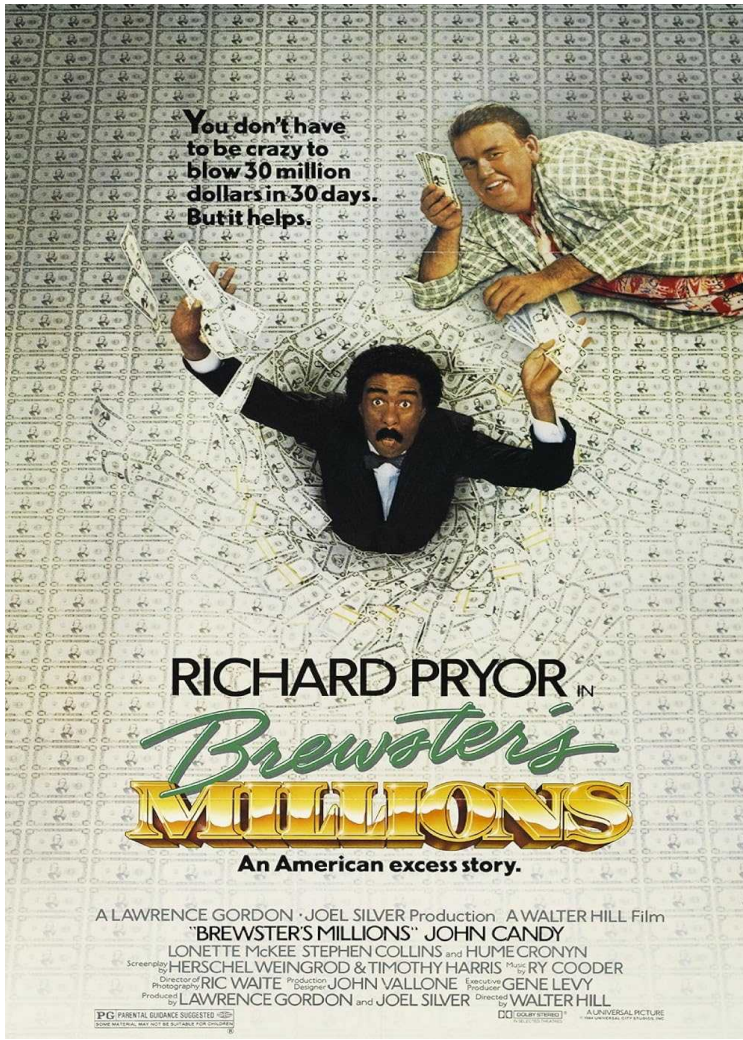
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

IQLA
International Quantitative Linguistics Association

False dichotomy!

- Exposition
- Rising action
- Falling action
- Discussion




You don't have to be crazy to blow 30 million dollars in 30 days. But it helps.
RICHARD PRYOR IN
Brewster's
MILLIONS
An American excess story.
A LAWRENCE GORDON · JOEL SILVER Production A WALTER HILL Film
"BREWSTER'S MILLIONS" JOHN CANDY
LONETTE McKEE STEPHEN COLLINS and HUME CRONYN
Screenplay by HERSCHEL WEINGROD & TIMOTHY HARRIS Music by RY COODER
Director of Photography RIC WAITE Production Designer JOHN VALLONE Executive Producer GENE LEVY
Produced by LAWRENCE GORDON and JOEL SILVER Directed by WALTER HILL
PG PARENTAL GUIDANCE SUGGESTED
A UNIVERSAL PICTURE

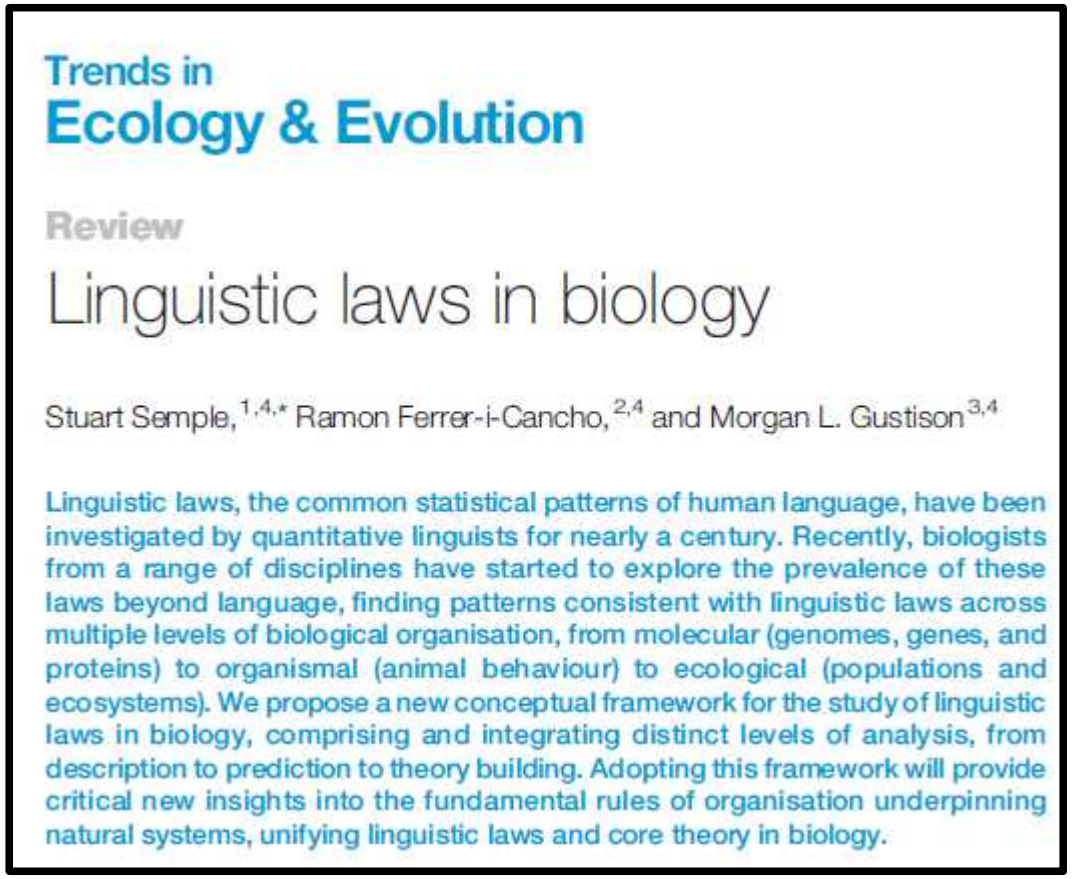
Where are we? Quantitative Linguistics

Discipline that investigates language and communication systems from a mathematical and physical perspective.

"Qualitative linguistics searches for rules and enumerates the exceptions, quantitative linguistics searches for models of phenomena and tests the models, just as in physics."
Gabriel Altmann (1931-2020), *Glottometrics* 33 (2016:32).

Linguistic features
Scientific principles → **Linguistic Laws**
Physical magnitudes

Linguistic Theory from Quantitative Linguistics



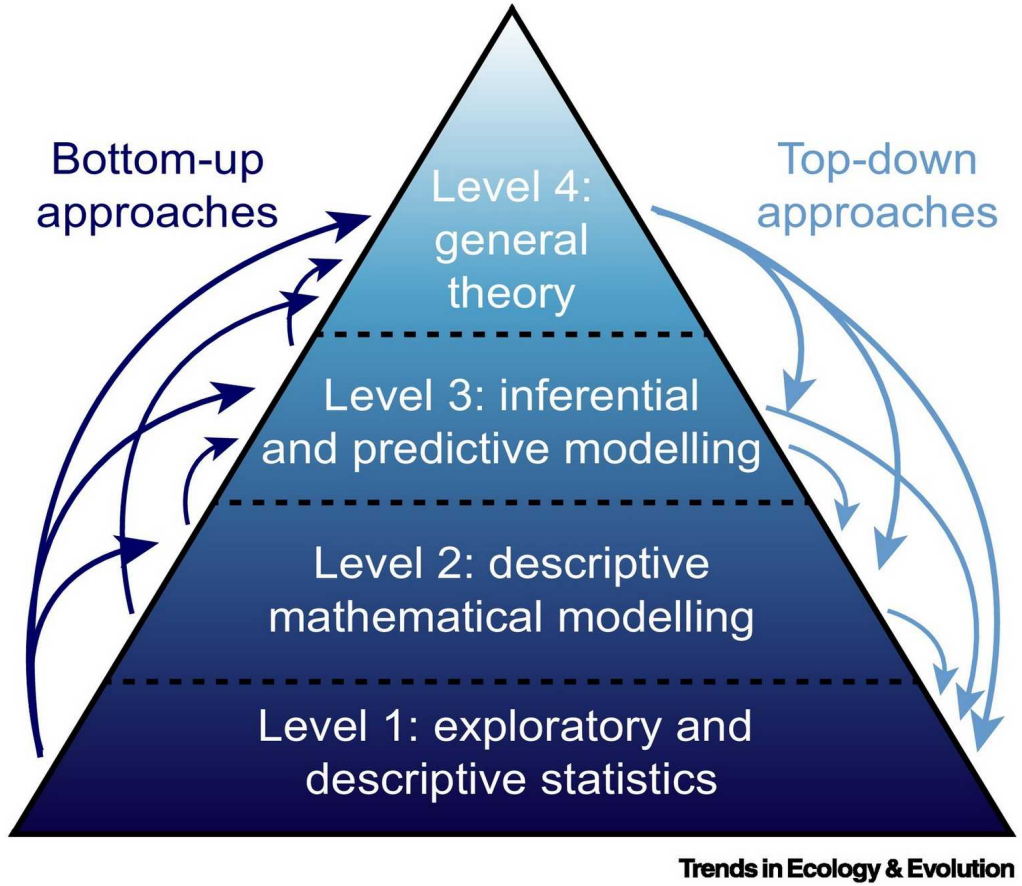
Trends in
Ecology & Evolution

Review

Linguistic laws in biology

Stuart Semple,[1,4,*] Ramon Ferrer-i-Cancho,[2,4] and Morgan L. Gustison[3,4]

Linguistic laws, the common statistical patterns of human language, have been investigated by quantitative linguists for nearly a century. Recently, biologists from a range of disciplines have started to explore the prevalence of these laws beyond language, finding patterns consistent with linguistic laws across multiple levels of biological organisation, from molecular (genomes, genes, and proteins) to organismal (animal behaviour) to ecological (populations and ecosystems). We propose a new conceptual framework for the study of linguistic laws in biology, comprising and integrating distinct levels of analysis, from description to prediction to theory building. Adopting this framework will provide critical new insights into the fundamental rules of organisation underpinning natural systems, unifying linguistic laws and core theory in biology.

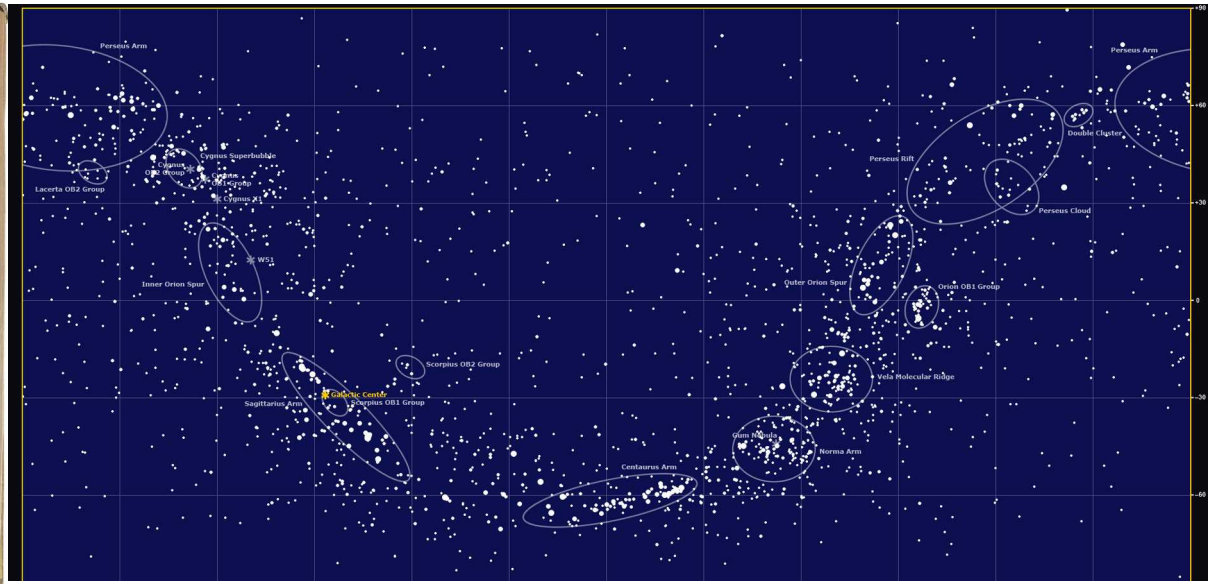

Bottom-up approaches
Top-down approaches
Level 4: general theory
Level 3: inferential and predictive modelling
Level 2: descriptive mathematical modelling
Level 1: exploratory and descriptive statistics
Trends in Ecology & Evolution

Semple, Ferrer-i-Cancho & Gustison (2022)

https://www.cell.com/trends/ecology-evolution/fulltext/S0169-5347(21)00230-5

Tycho Brahe & Johannes Kepler: Astronomy example (I)

EFFIGIES TYCHONIS BRAHE OTTONI DANÆ ÆTATIS SVÆ ANNO 50 COMPLETO.
QVO POST DIVTINVM IN PATRIA EXILIVM LIBERTATI DESIDERATÆ DIVINO PROVISV RESTITVTVS EST.

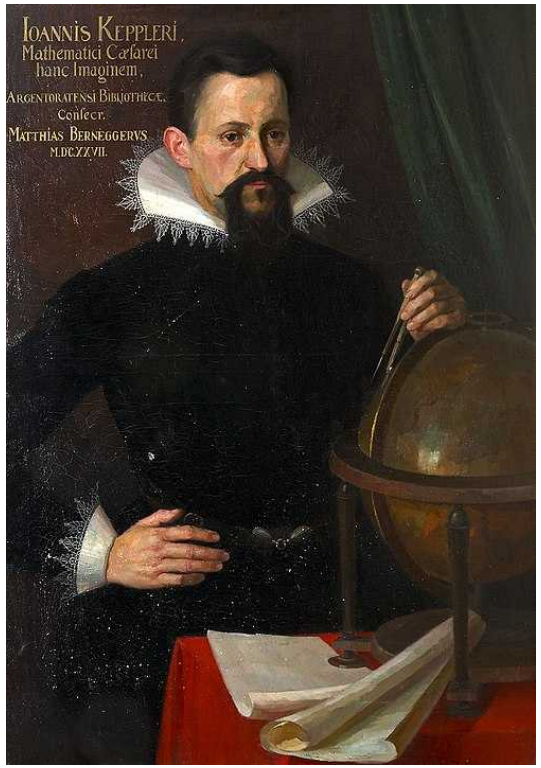


Tycho Brahe's Star Catalog

Tycho Brahe (1546-1601) life's work in astronomy consisted of measuring the positions of the stars, planets, Moon, and Sun, and carefully recording these measurements. → PREDICTIONS!

https://www.aanda.org/articles/aa/pdf/2010/08/aa14002-10.pdf

Tycho Brahe & Johannes Kepler: Astronomy example (II)

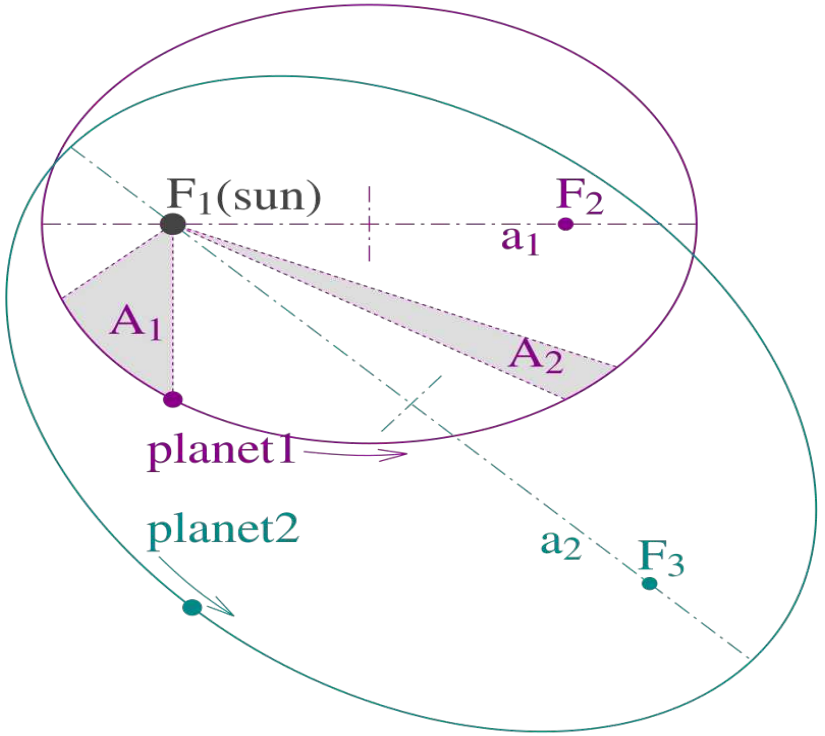


Johannes Kepler (1571-1630) deduces after analyzing Tycho Brahe's Catalogue the three principles known today as **Kepler's laws of planetary motion:**

1.- The orbit of a planet is an ellipse with the Sun at one of the two foci.

2.- A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.

3.- The square of a planet's orbital period is proportional to the cube of the length of the semi-major axis of its orbit.

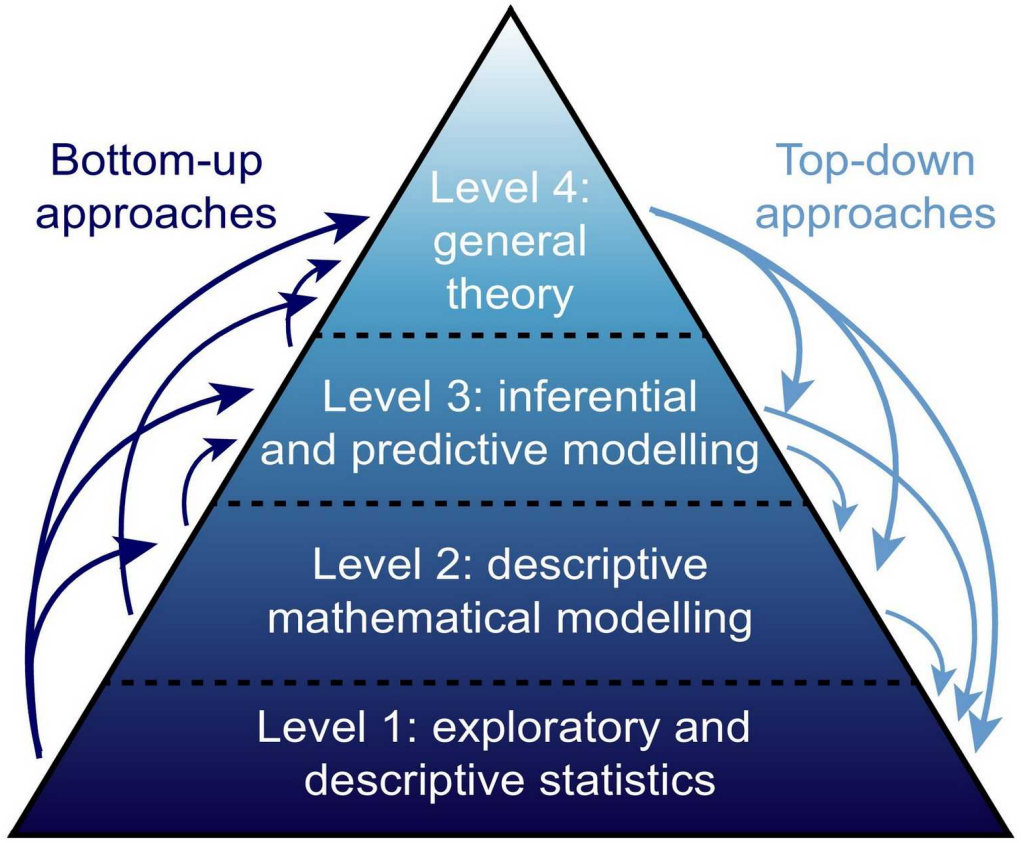


They are empirical laws, but they are not Newton's laws of universal gravitation!

TO PREDICT vs TO EXPLAIN!

https://en.wikipedia.org/wiki/Johannes_Kepler

Linguistic Theory from Quantitative Linguistics


Bottom-up approaches
Top-down approaches
Level 4: general theory
Level 3: inferential and predictive modelling
Level 2: descriptive mathematical modelling
Level 1: exploratory and descriptive statistics
Trends in Ecology & Evolution

Box 1. Key methodological issues in the study of linguistic laws

Two main statistical approaches are used to investigate linguistic laws: model fitting and correlation. Model fitting, which is the traditional approach taken in quantitative linguistics, involves finding a mathematical function (i.e., an equation) that defines the dependency between two variables. Putative mathematical functions for a number of linguistic laws are shown in Table 1 in main text. A key advantage of the model fitting approach is that it helps characterise the nonlinear behaviour of a law. Any resulting equations are then tailored by modifying parameters to fit new datasets (e.g., Zipf's rank-frequency exponent varies across languages [35]). However, these fitted equations are only first approximations to the true, highly complex, relationship between variables [34,93,94]. The validity of many of these equations has been questioned using both theoretical and empirical arguments [32,95–97]. As a reaction to these challenges, as well as a need to generalise laws to diverse systems [84], correlational approaches have become increasingly popular (e.g., [45]).

A correlational approach involves the use of statistical tools to test for positive or negative associations between variables. This approach has some key advantages. First, it simplifies the analysis with respect to curve fitting, and the use of appropriate statistical tools (e.g., Spearman's correlation, linear mixed effect models) removes some assumptions about how variables are structured and related to one another, or controls for the effect of multiple factors. Second, it is supported by theoretical arguments involving optimal coding that predict a negative (or null) correlation consistent with Zipf's law of abbreviation or Menzerath's law [85]. A limitation of correlational analysis is that it is not appropriate for all laws, specifically those where variable definitions are intrinsically related (e.g., frequency can only decrease as rank increases in Zipf's rank-frequency law; type number can only increase with token number in Herdan's law; Table 1 in main text). For other laws, it is important to confirm that there are not 'inevitable' correlations due to variables being functionally dependent [79,98]. Researchers have used several strategies to successfully address this criticism for Menzerath's law and Zipf's law of abbreviation [23,26,99].

Semple, Ferrer-i-Cancho & Gustison (2022)

https://www.cell.com/trends/ecology-evolution/fulltext/S0169-5347(21)00230-5

Level 1: LLM's in exploratory and descriptive statistics

Relevance of LLMs:

LLMs can detect statistical regularities in language that mirror known linguistic laws (e.g., Zipf's law, Menzerath's law...), since their training involves analyzing massive corpora.

They can assist in automated corpus analysis across modalities (text, code, biosequences), enabling high-throughput screening for patterns consistent with linguistic laws **in both language and biological systems.**

LLMs may uncover new candidate laws or exceptions via unsupervised anomaly detection or clustering methods.

Level 1: LLM in exploratory and descriptive statistics

Epistemological note:

a) **Their use raises the challenge of data-dependence**: patterns may reflect corpus composition (**bias**) rather than fundamental laws.
b)The black-box nature of LLMs complicates tracing whether a given regularity arises from learned linguistic laws or training data artifacts.

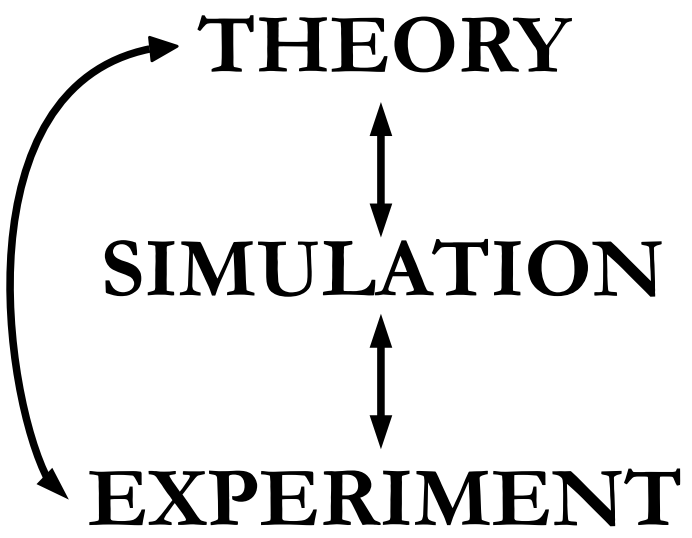
Applications:
a) Used to generate synthetic biological datasets for test linguistic laws.
b) Employed in cross-lingual or cross-species comparisons by modeling behavioral or genomic data as language-like sequences.

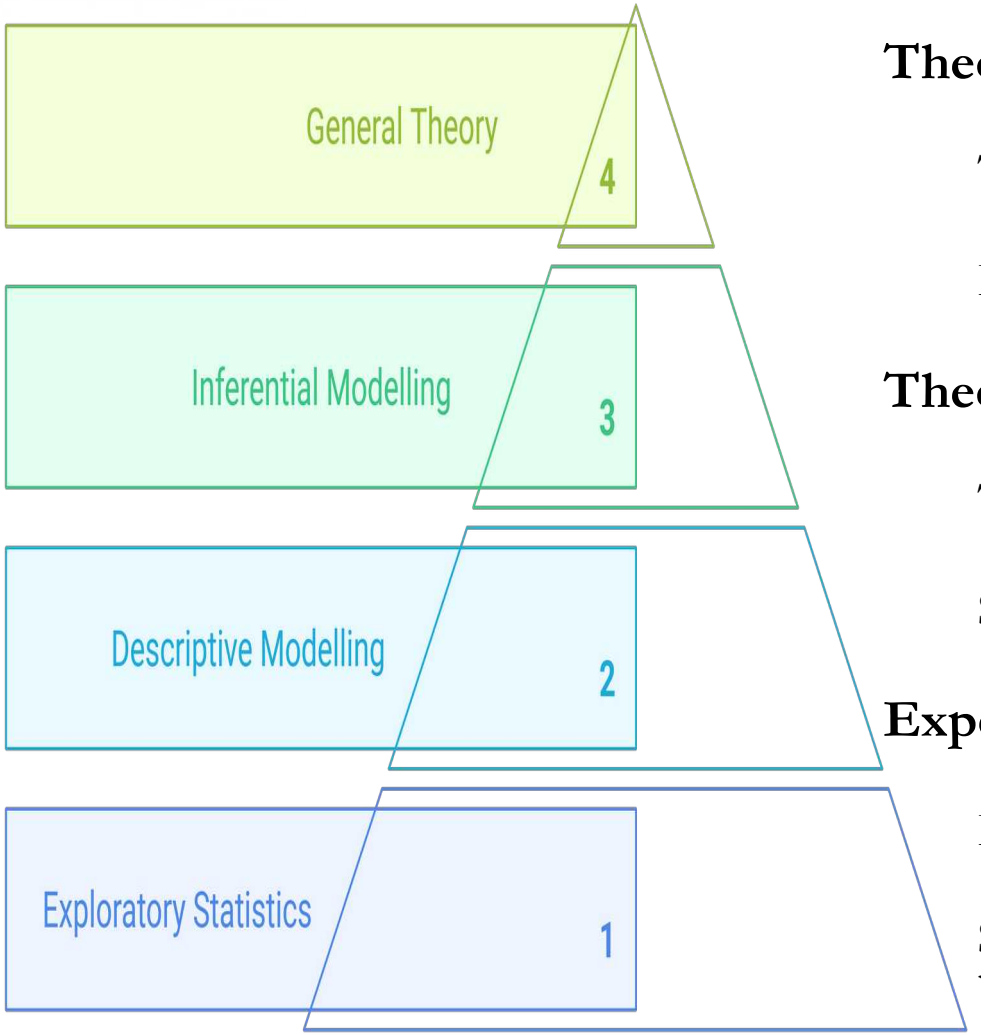
Level 2: Descriptive Mathematical Modelling

Relevance of LLMs:

→ LLMs are not inherently mathematical models, but they can perfectly simulate outputs consistent with linguistic laws.

→ LLMs can be used to generate data distributions which are then tested against theoretical models (e.g., do LLMs' outputs follow Zipfian distributions? Lognormality in data?).





Theory ⇄ Experiment

Theories guide experiments (→)

Experimental results validate or refine theories (←)

Theory ⇄ Simulation (LLM)

Theories provide the foundational models for simulations (→)

Simulations can generate new insights that refine theories (←)

Experiment ⇄ Simulation (LLM)

Experimental data is used to calibrate and validate simulations (→)

Simulations can help design experiments or predict outcomes before real-world testing (←).

Modified from Semple, Ferrer-i-Cancho & Gustison (2022)

Level 2: Descriptive Mathematical Modelling

Epistemological note:

a) LLMs blur the line between empirical data and model: **they are both a product of language and a generator of language-like phenomena.**
b) Their ability to emulate/simulate/predict known patterns may be mistaken for understanding, **necessitating careful interpretation.**

Applications:

a) Support for **simulation** and exploratory modelling of systems not easily observable (e.g., extinct (impossible?) languages, hypothetical proteomes...).
b) Help frame null models or synthetic baselines in comparative studies.

Level 3: Inferential and Predictive Modelling

Relevance of LLMs:

LLMs, particularly when coupled with additional architecture (e.g., reinforcement learning, symbolic reasoning), can be harnessed **to develop and test hypotheses about the function of linguistic laws in communication systems.**

They support **simulation-based inference** (e.g., via ablation studies or prompt engineering) to test the impact of compression, redundancy or other information theory principles in communicative systems.

Level 3: Inferential and Predictive Modelling

Epistemological note:

a) The **predictive power of LLMs does not entail explanatory adequacy:** **LLM may predict effects without modelling or explain causes.**
b) There's a clear **risk of overfitting to shallow regularities**, especially if not constrained by biologically meaningful priors.

Applications:
a) Predicting how biological communication systems might evolve under constraints of efficiency, or other principles.
b) Developing plausible theorethical models that test functional hypotheses (e.g., compression as an adaptive trait).

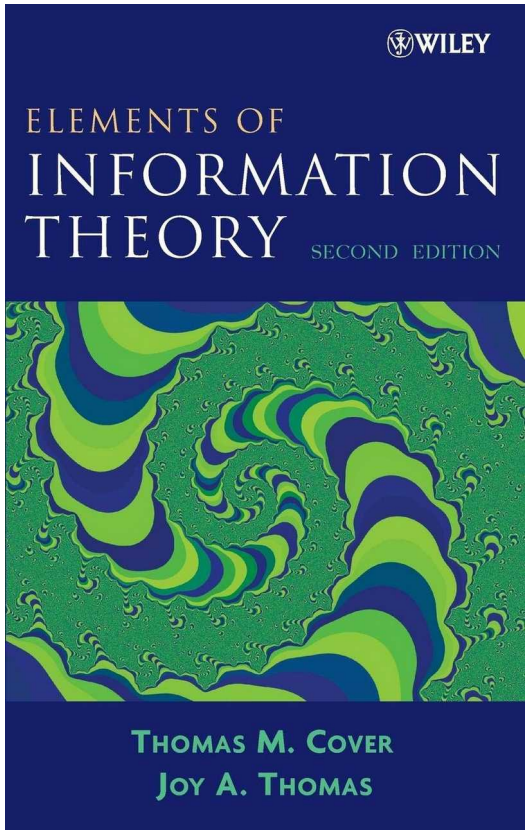
Level 4: General Linguistic Theory

Relevance of LLMs:

LLMs **contribute to reframing theories of language as emergent from optimization under constraints** —a central idea also proposed in Semple et al. (2022).

LLMs **provide new empirical support** for compression, efficiency, and information trade-off principles.

LLMs involve reconsidering the nature of artificial communication, **technosemiotics vs biosemiotics.**



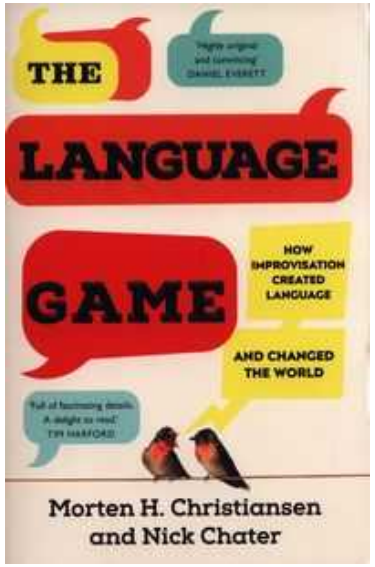
Hernández-Fernández (2021) https://upcommons.upc.edu/handle/2117/346191

Level 4: General Linguistic Theory

Epistemological note:

a) **Do LLMs force us to redefine theories of meaning, representation, and intention—all central to linguistic theory?**

b) LLMs challenge traditional views by showing that **statistical learning alone can lead to high-level language use**, raising questions about what constitutes understanding or rule-following.



Applications:

a) LLMs serve as **computational analogs** for general theories of cognition, communication, and organization in complex systems.

b) **Inspire cross-domain models** (e.g., energy allocation as a general principle across language and communication systems).

Final examples

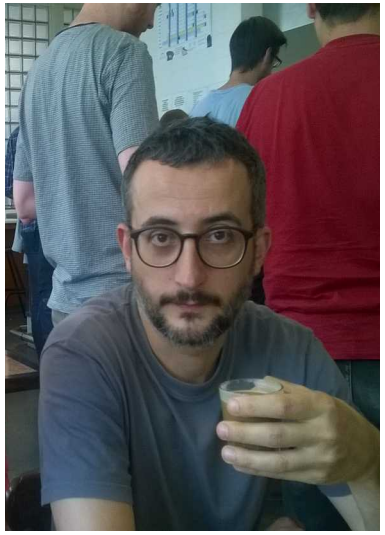
Level	LLM Contribution	Challenge	Example Use
1. Descriptive	Pattern detection	Data bias	Zipfian analysis on biological data
2. Mathematical modelling	Simulation of linguistic laws	Model-data conflation	Testing Menzerath-like structures in output
3. Predictive	Hypothesis testing	Predict ≠ explain	Exploring cost-compression trade-offs
4. Theoretical	Theory generation	Redefining meaning	Integration of Statistical Learning in Linguistic Theories

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

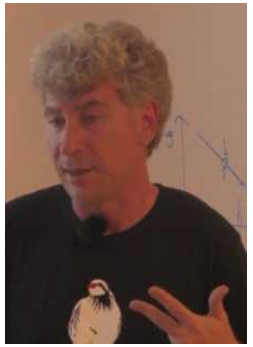
Thank you!
Moltes Gràcies!
Moc děkuji!



Jaume Baixeries Juvillà (UPC)

Neus Català Roig (UPC)

Bernardino Casas (UPC)



Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA International Quantitative Linguistics Association

Quantitative Linguistics and Large Language Models: Epistemological Challenges and Applications

Open Discussion and Debate

QUALICO 2025
Brno, Czech Republic
June 26-28, 2025

antonio.hernandez@upc.edu @therfer

References

Català, N., Casas, B., Hernández-Fernández, A. (2023). The semanticity of catalan words: quantitative linguistics in the era of large language models. In: "Tejiendo palabras: explorando la lengua, la lingüística y el proceso de traducción en la era de la inteligencia artificial". Madrid: Dykinson, 2024, p. 249-268. http://hdl.handle.net/2117/413308

Institut d'Estudis Catalans

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

IQLA
International Quantitative Linguistics Association

EXTRA MATERIAL

LLM vs QL models (I)

Aspect of Linguistics	Large Language Models (LLMs)	Quantitative Linguistics
Data Processing	It needs and processes vast amounts of data to learn and generalize linguistic patterns and structures through learning algorithms.	Focuses on statistical analysis of linguistic data, employing quantitative methods to derive patterns and linguistic laws, requiring less data than LLMs.
Approach to Language	Understands language through contextual analysis, capturing nuanced semantic and syntactic relationships for generating human-like text.	Emphasizes statistical probabilities and mathematical models and methods to analyze linguistic elements, often without considering contextual nuances.

Català et al. (2023)

LLM vs QL models (II)

Flexibility and Adaptability	Demonstrates adaptability in learning from diverse or even rare linguistic sources and adjusting to various styles and contexts for text generation.	It is based on statistical models and linguistic rules or patterns observed in human language, and therefore offers less flexibility to adapt to various unusual communicative styles or contexts.
Processing Context	Proficiency in processing and using contextual information to produce coherent and contextually relevant texts. Beyond a parallel with neural networks, there is no concern to establish connections with biology.	Focuses more on statistical correlations among linguistic elements, sometimes lacking in-depth comprehension of nuanced contextual information, but seeking to contextualize linguistic laws within the sciences, especially physics and biology.

Català et al. (2023)

LLM vs QL models (III)

Flexibility and Adaptability	Demonstrates adaptability in learning from diverse or even rare linguistic sources and adjusting to various styles and contexts for text generation.	It is based on statistical models and linguistic rules or patterns observed in human language, and therefore offers less flexibility to adapt to various unusual communicative styles or contexts.
Processing Context	Proficiency in processing and using contextual information to produce coherent and contextually relevant texts. Beyond a parallel with neural networks, there is no concern to establish connections with biology.	Focuses more on statistical correlations among linguistic elements, sometimes lacking in-depth comprehension of nuanced contextual information, but seeking to contextualize linguistic laws within the sciences, especially physics and biology.

Català et al. (2023)

LLM vs QL models (IV)

Empirical Validation	Lack direct physical or experimental validation due to reliance on learned patterns and vast parameters, limiting direct empirical verification.	Focuses on empirical validation of linguistic theories by testing mathematical models against observed linguistic data, aiming for falsifiable predictions and experimental verification.
Occam's Razor Application	Generally, not bound by the principle of simplicity or parsimony, as the focus is on the effectiveness of learning from extensive data, leading to complex models with a myriad of parameters.	Applies Occam's Razor to strive for simpler linguistic models that explain observed linguistic phenomena without unnecessary complexity, favoring elegance and explanatory power.
Interpretation of Data	Prioritizes effective modeling and generation of language without rigid adherence to physical, linguistic or experimental laws or interpretation of linguistic data.	Emphasizes interpreting linguistic data through the lens of mathematical-physical models, aiming to uncover underlying linguistic principles and patterns for scientific understanding.

Català et al. (2023)

LLM vs QL models (V)

Role in Linguistic Theory	Offers a powerful tool for language generation and processing but may lack direct support for theoretical linguistic principles without empirical grounding.	Provides a framework for developing and testing linguistic theories, striving for empirical support and scientific validation of proposed linguistic principles.
Technoethics	LLM training has a huge energy consumption and environmental impact, as well as resorting to labour exploitation of data taggers from impoverished countries. It also has a data protection risk and a social and employment impact in many sectors.	Has a much lower energy and computational cost than LLMs and there is no labour exploitation of researchers in the discipline, beyond the general working conditions of fellows and doctoral researchers.

Català et al. (2023)